A Comparative Study of Data Mining Techniques for Student Performance Analysis in Education

Zahira Noor Quraishi[1], Dr. Atul Dattarya Newase[2]

[1]Research Scholar, [2]Research Supervisor
[1,2]Dr. A. P. J. Abdul Kalam University, Indore, India
zahira16sep@gmail.com , dr.atulnewase@gmail.com

Presented at International Conference on Engineering, Economics, Management and Applied Sciences (ICE2MAS-24), Bangkok, 21-24 December 2024, organized by Academy of Art, Science and Technology (AAST).

Published in IJIRMPS (E-ISSN: 2349-7300), ICE2MAS-24

License: Creative Commons Attribution-ShareAlike 4.0 International License

Abstract

The rapid growth of digital technologies in education has led to the generation of large volumes of data related to student learning, assessment, and behavior. Educational Data Mining (EDM) has emerged as a powerful approach for analyzing such data to improve academic decision-making and student outcomes. This study presents a comparative analysis of data mining techniques used for student performance analysis across different educational contexts. Based on a synthesis of recent empirical studies and systematic literature reviews, the research examines commonly applied methods including decision trees, random forest, naïve Bayes, neural networks, clustering techniques, feature selection approaches, and process frameworks such as CRISP-DM. The comparison highlights variations in model performance, interpretability, scalability, and applicability across higher education, secondary education, and mixed learning environments. Findings indicate that ensemble models, particularly random forest, consistently achieve strong predictive accuracy, while simpler models offer better transparency for educational stakeholders. However, the analysis also reveals limitations such as dataset dependency, lack of generalizability, data quality issues, and underrepresentation of school-level education in empirical research. By identifying strengths, weaknesses, and methodological gaps, this study provides guidance for selecting appropriate data mining techniques and supports the development of more robust, explainable, and context-aware student performance analysis systems.

Keywords: Educational Data Mining, Student Performance Analysis, Machine Learning, Data Mining Techniques, Learning Analytics

1. Introduction

Education systems worldwide are undergoing digital transformation driven by the widespread adoption of Information and Communication Technologies (ICT), online learning platforms, and data-driven management tools. These developments have resulted in the continuous generation of large and complex educational datasets, including academic records, attendance logs, learning management system data, assessment results, and feedback from students and teachers. Effectively analyzing this data has become

essential for understanding learning behavior, improving teaching strategies, and enhancing overall educational quality [1].

Educational Data Mining (EDM) is a research field that applies data mining, machine learning, and statistical techniques to extract meaningful patterns from educational data. One of the most prominent applications of EDM is student performance analysis, which aims to predict academic outcomes, identify at-risk learners, and support early interventions. Accurate performance analysis enables educators and administrators to personalize instruction, optimize resource allocation, and improve retention and completion rates [2].

A wide range of data mining techniques has been employed for student performance analysis. Traditional machine learning models such as decision trees, naïve Bayes, linear and logistic regression, and support vector machines are widely used due to their simplicity and interpretability. Ensemble methods, particularly random forest, have gained popularity for their ability to handle high-dimensional data and improve predictive accuracy. In addition, clustering techniques such as K-means are used to group students based on learning behavior and performance patterns, while feature selection methods enhance model efficiency by identifying the most influential attributes [3].

Recent studies have also explored advanced approaches, including neural networks, fuzzy-based models, optimized algorithms, and process-oriented frameworks such as CRISP-DM. These methods aim to address challenges such as uncertainty in human decision-making, complex learning behaviors, and heterogeneous educational environments. Despite these advancements, the literature reveals that many studies are limited to specific contexts, often focusing on higher education institutions or single datasets. This raises concerns regarding the generalizability and scalability of proposed models across diverse educational systems [4].

Another key issue highlighted in the literature is data quality. Missing values, imbalanced class distributions, and heterogeneous data sources can significantly affect model performance. While techniques such as resampling, preprocessing, and feature engineering are employed to mitigate these issues, there is no standardized framework for addressing them across studies. Furthermore, although high predictive accuracy is frequently reported, interpretability and practical usability remain critical concerns, particularly for educators and policymakers who require transparent and actionable insights [5].

Comparative studies play an important role in addressing these challenges by evaluating the strengths and limitations of different data mining techniques under varied conditions. Such comparisons help identify models that balance accuracy, interpretability, and computational efficiency. However, existing comparative analyses are fragmented and often restricted to limited datasets or specific educational levels [6].

In this context, the present study aims to provide a comprehensive comparative analysis of data mining techniques used for student performance analysis in education. By synthesizing findings from recent empirical research and systematic reviews, the study examines commonly applied methods, highlights methodological trends, and identifies existing research gaps. The objective is to support informed model selection and encourage the development of robust, explainable, and context-sensitive EDM solutions that can effectively enhance student performance across educational settings.

2. Literature review

Ordonez-Avila et al. (2023), Teacher evaluation in higher education is an important research area that relies on heterogeneous data from multiple academic stakeholders, particularly students, who provide rich and meaningful feedback. This study conducts a systematic literature review to identify research on predicting teacher evaluation using student performance data. Following structured phases of planning, selection, and extraction, the review highlights extensive use of educational data mining techniques, including fuzzy-based models, neural networks, decision trees, and machine learning classifiers. The

findings indicate growing emphasis on fuzzy principles, reflecting the inherent uncertainty of human decision-making in evaluation processes. **[1]**

Ashish et al. (2024), The integration of Information and Communication Technology (ICT) has significantly transformed modern educational environments, enabling interactive, adaptive, and learner-centered teaching approaches. Digital platforms, virtual classrooms, and multimedia tools have reshaped traditional pedagogy and expanded opportunities to enhance student performance. This study investigates the role of ICT in improving student learnability by applying data mining techniques to educational datasets. By analyzing patterns and correlations between ICT usage and academic outcomes, the research aims to provide evidence-based insights that support effective technology-enhanced learning and inform future educational practices. **[2]**

Staneviciene et al. (2024), This study examines the use of data analytics and machine learning to predict academic outcomes and support sustainable e-learning, particularly in university technological programs with high dropout rates. Student Performance Prediction (SPP) is emphasized for its ability to enable personalized learning and early interventions. Using a case study guided by the CRISP-DM methodology, the research applies classification, regression, and clustering techniques to academic data. Despite challenges related to data quality and completeness, results demonstrate that mining learning process data can effectively identify at-risk students and support educational quality aligned with Sustainable Development Goal 4. **[3]**

Kaur et al. (2023), Educational Data Mining (EDM) and Learning Analytics are increasingly used to address critical educational challenges such as student performance prediction and assessment. This systematic literature review analyzes studies published between 2012 and 2022 to examine techniques employed in both domains. EDM focuses on applying data mining methods to educational data, while learning analytics emphasizes understanding learning processes through visualization and quantitative analysis. Reviewing 41 relevant studies, the paper identifies commonly used statistical, machine learning, and visualization techniques, highlights research gaps, and provides guidance for educators, researchers, and policymakers seeking to enhance learning outcomes through data-driven approaches. **[4]**

Missah et al. (2023), The application of Data Mining (DM) in education supports evidence-based decision-making by educational leaders. This review examines how DM research is distributed across different educational levels, with a particular focus on Basic Education (BE). Analyzing 94 studies published between 2017 and 2022 from nine major academic publishers, the findings reveal a strong imbalance: most EDM research targets tertiary education, while basic and pre-tertiary levels remain underrepresented. Additionally, existing studies focus heavily on student performance factors, overlooking critical aspects such as pedagogical resources, revealing both population and knowledge gaps in EDM research. **[5]**

Sun et al. (2024), Data mining techniques in EDM are widely used to analyze student preferences and learning behaviors, particularly in e-learning environments. This study proposes a novel model, the Chaotic-Tuned Shuffled Frog Leaping Optimized Random Forest (CSFLO-RF), to improve learning behavior prediction. Student activity logs are clustered using K-Means to identify behavior patterns and learning styles, which are then classified using the proposed model. Experimental results, implemented on the WEKA platform, demonstrate that CSFLO-RF outperforms existing methods, highlighting its effectiveness in forecasting student behavior and enhancing adaptive learning systems. **[6]**

Abideen et al. (2023), Student enrollment analysis is a critical but underexplored area in educational data mining, particularly in developing regions. This study focuses on predicting school enrollment in Punjab, Pakistan, using five years of data from 100 schools. Machine learning algorithms—Multiple Linear Regression, Random Forest, and Decision Tree—are applied to identify significant features and predict future enrollment trends. The proposed model supports enrollment target classification and provides insights to address low enrollment levels, thereby contributing to improved planning, literacy rates, and more effective education policy implementation. **[7]**

Wang et al. (2024), In secondary education, large volumes of student achievement data are often underutilized, limiting their potential to support educational improvement. This study highlights how data mining techniques can enhance education informatization by analyzing secondary school performance data more effectively. Traditional teaching methods are identified as insufficient for engaging students, prompting the adoption of case-based teaching supported by data analysis. By statistically evaluating student performance and applying targeted instructional strategies, the study demonstrates how data mining can improve learning outcomes and support the broader goals of quality and vocational education. **[8]**

Syed Mustapha et al. (2023), Accurate prediction of academic success increasingly depends on effective feature engineering and selection in machine learning models. This study compares feature selection methods—Boruta and Lasso for regression, and Recursive Feature Elimination (RFE) and Random Forest Importance (RFI) for classification—using the OULA dataset. Results show that Gradient Boost with Boruta achieved the lowest prediction error, while RFI produced the highest classification accuracy. The findings emphasize that appropriate feature selection significantly improves model performance, offering valuable guidance for developing reliable student success prediction systems. **[9]**

Alghamdi et al. (2023), Predicting academic performance in early secondary education can help institutions identify at-risk students and provide timely support. This study uses data from high school graduates in the Al-Baha region of Saudi Arabia to develop predictive models using Naïve Bayes, Random Forest, and J48 algorithms. Data balancing with SMOTE and feature extraction using correlation coefficients improved model reliability. Performance evaluation through cross-validation showed exceptionally high accuracy, with Naïve Bayes achieving 99.34%, demonstrating the effectiveness of EDM techniques in early academic performance prediction. **[10]**

Chan et al. (2023), Despite the rapid growth of Educational Data Mining (EDM), its application in secondary school contexts remains limited. This literature review analyzes 18 studies published between 2008 and 2021 focusing on secondary school data. Most studies address academic success classification, influence factor analysis, and dropout risk prediction using traditional machine learning methods. Advanced techniques such as deep learning and knowledge tracing are rarely applied. The review identifies a clear research gap and emphasizes the need to expand EDM research into secondary education to better support learning and policy decisions. **[11]**

Collier et al. (2024), This article bridges Educational Data Mining (EDM) with Research Methods, Measurement, and Evaluation (RMME), introducing RMME researchers to EDM's goals and analytical culture. While RMME typically prioritizes parameter estimation and statistical inference, the paper emphasizes EDM's use of statistics and machine learning to develop practical, high-performing methods for educational contexts. It addresses three guiding questions: the main interests of each community, their discipline-specific vocabulary, and how their approaches to similar data differ or overlap. By clarifying shared ground and distinctions, the paper supports more effective cross-disciplinary communication and collaboration. **[12]**

Assiri et al. (2024), Saudi Arabian university admissions often rely on cumulative scores that may not fit all majors, contributing to failure, dropout, and transfers. This study analyzes relationships between university GPA and admission features using data mining, proposing a Jaccard-based similarity model (including modified variants) plus distribution and correlation analyses. Findings show admission-performance relationships vary by major, revealing weaknesses in one-size-fits-all policies and emphasizing hidden details like high school course grades. Machine learning models then classify students into suitable majors; KNN achieved 100% accuracy, outperforming decision tree and SVM, supporting improved major placement aligned with skills and interests. **[13]**

Huerta et al. (2023), This work applies the Knowledge Discovery in Databases (KDD) methodology using RapidMiner to filter and organize information for more efficient decision-making, with a focus on

historical investment per student in the education sector. By mining patterns from stored data, the study aims to reduce waste caused by poor information management and support more accurate prediction. Findings indicate that expenditure per student generally increases over time, though allocation differs by province, while still showing an upward trend overall. The study concludes that KDD-based analysis can visualize spending variation across education grades and provide relevant insights for future research and planning. **[14]**

Table 1. Systematic literature review

Ref	Author (First author et al.)	Year	Title	Methods	Result	Advantage	Limitation
[1]	Ordonez-Avila et al.	2023	Data mining techniques for predicting teacher evaluation in higher education: A systematic literature review	Systematic literature review; fuzzy models, neural networks, decision trees, ML classifiers	Identified growing use of fuzzy-based EDM for teacher evaluation prediction	Addresses uncertainty in human evaluation; comprehensive synthesis	Focused only on higher education; no empirical validation
[2]	Ashish et al.	2024	Transforming Education: The Impact of ICT and Data Mining on Student Outcomes	Data mining analysis of ICT usage and academic performance	Demonstrated positive correlation between ICT integration and student learnability	Supports evidence-based ICT adoption	Lacks detailed model comparison and validation
[3]	Staneviciene et al.	2024	Data mining-based prediction of students' performance for sustainable e-learning	CRISP-DM; classification, regression, clustering	Successfully identified at-risk students despite data challenges	Supports SDG 4; practical case study	Data quality and completeness issues
[4]	Kaur et al.	2023	Role of educational data mining and learning analytics techniques used for predictive modeling	Systematic literature review (2012–2022)	Identified common EDM and LA techniques and research gaps	Clear comparison of EDM vs learning analytics	Limited to reviewed studies; no experimentation
[5]	Missah et al.	2023	Systematic Review of Data Mining in Education on the Levels and Aspects of Education	Review of 94 studies (2017–2022)	Revealed imbalance toward tertiary education	Highlights population and knowledge gaps	Does not propose predictive models
[6]	Sun et al.	2024	Educational Technology Based on Data Mining: Mining Student Behavior Patterns	K-Means clustering; CSFLO-optimized Random Forest	CSFLO-RF outperformed existing models in behavior prediction	Improves adaptive learning systems	Model complexity may limit scalability
[7]	Abideen et al.	2023	Analysis of enrollment criteria in secondary schools using machine learning	Linear Regression, Random Forest, Decision Tree	Accurate enrollment trend prediction	Supports planning and education policy	Region-specific (Punjab, Pakistan)
[8]	Wang et al.	2024	Statistical Analysis of Secondary School Achievement Based on Data Mining	Statistical analysis; case-based teaching strategies	Improved student engagement and learning outcomes	Practical classroom application	Limited methodological detail
[9]	Syed Mustapha et al.	2023	Predictive analysis of students' learning performance	Feature selection (Boruta, Lasso, RFE, RFI); ML models	Feature selection significantly improved accuracy	Strong guidance on feature engineering	Tested on a single dataset (OULA)

[10]	Alghamdi et al.	2023	Data mining approach to predict success of secondary school students	Naïve Bayes, Random Forest, J48; SMOTE	Naïve Bayes achieved 99.34% accuracy	Effective early identification of at-risk students	Possible overfitting; regional dataset
[11]	Chan et al.	2023	A literature review on educational data mining with secondary school data	Literature review (2008–2021)	Identified lack of advanced methods in secondary EDM	Clearly defines research gaps	Limited number of reviewed studies
[12]	Collier et al.	2024	Discovering educational data mining: An introduction	Conceptual comparison of EDM and RMME	Clarified methodological overlaps and differences	Encourages interdisciplinary collaboration	Not an empirical EDM study
[13]	Assiri et al.	2024	Enhanced student admission procedures using data mining	Jaccard similarity, KNN, DT, SVM	KNN achieved 100% accuracy in major placement	Improves admission decision accuracy	Admission-focused; not post-admission performance
[14]	Huerta et al.	2023	Data mining: Application of digital marketing in education	KDD methodology using RapidMiner	Identified trends in investment per student	Supports policy and planning decisions	Descriptive, not predictive

3. Research Gap

A clear research gap emerging from the above studies is that, although Educational Data Mining (EDM) is widely applied for student performance prediction, behavior mining, admissions, and teacher evaluation, there is limited integration of these approaches into unified, school-focused and decision-actionable frameworks, especially beyond tertiary contexts. Several works are systematic reviews that map techniques (e.g., fuzzy models, ML classifiers, visualization, and learning analytics) and highlight research gaps, but they do not provide deployable models or cross-context validation for real educational settings [1], [4], [5], [11]. Empirical studies demonstrate strong performance of specific models (e.g., optimized Random Forest for behavior prediction, feature selection improving accuracy, or very high accuracy in secondary performance prediction), yet they are often constrained by single datasets, localized samples, or context-specific conditions, raising concerns about generalizability, reproducibility, and potential overfitting [6], [9], [10]. In addition, key practical challenges such as data quality, missing data, and heterogeneous stakeholder-driven inputs (e.g., teacher evaluations) are recognized but remain insufficiently addressed through standardized pipelines and robust validation strategies [1], [3]. Finally, while ICT adoption and data-driven teaching strategies are discussed as beneficial, existing studies provide limited comparative evidence linking technology-use indicators to interpretable predictive features and scalable interventions across diverse school environments [2], [8].

4. Systematic result analysis

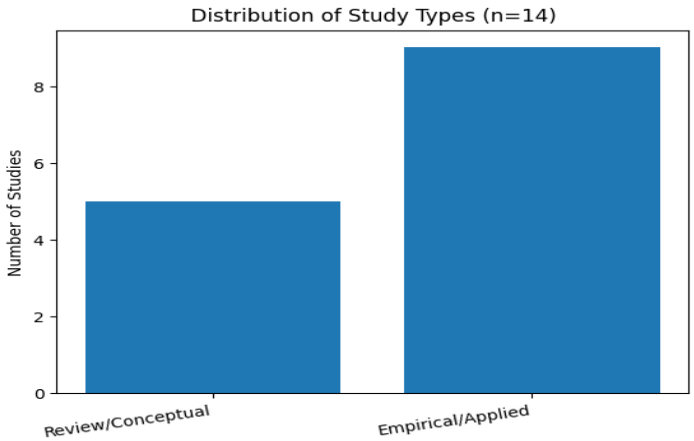

Figure 1: Distribution of Study Types (n=14)

Figure 1 shows the distribution of studies by research type. The literature is dominated by empirical/applied studies, while a smaller portion consists of review/conceptual papers. This indicates that, alongside synthesis work, many researchers are actively implementing EDM models, but systematic consolidation and cross-validation efforts are comparatively fewer.

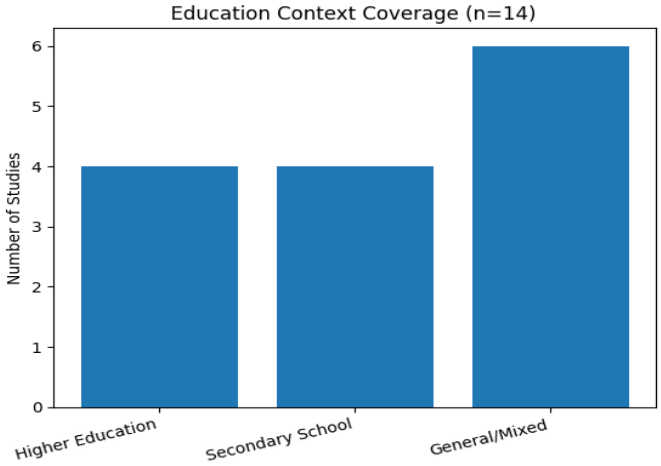

Figure 2: Education Context Coverage (n=14)

Figure 2 summarizes the education contexts addressed by the reviewed studies. A larger share of the work is categorized as general/mixed education, while higher education and secondary school contexts appear equally represented. This suggests that although EDM is broadly discussed across education, context-specific evidence for school education (especially secondary and earlier levels) is still limited in depth compared with university-centered applications.

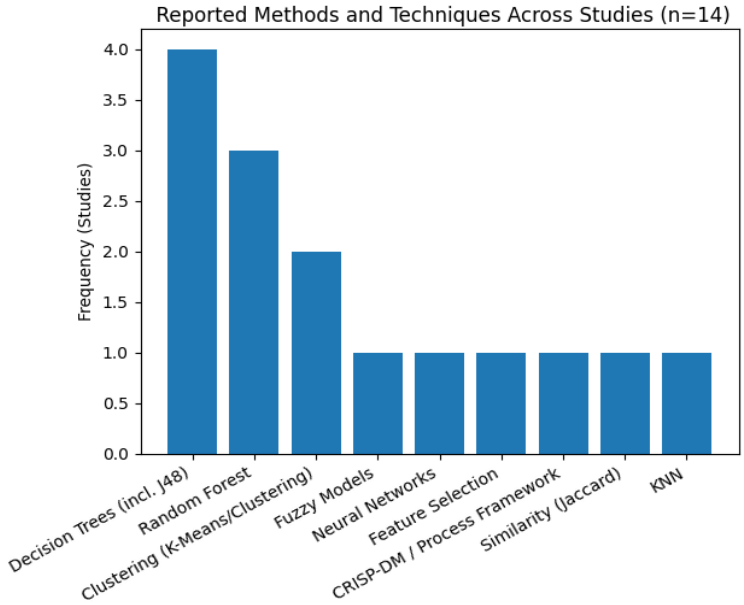

Figure 3: Reported Methods and Techniques Across Studies (n=14)

Figure 3 presents the frequency of methods explicitly reported in the studies. Decision-tree-based approaches are the most mentioned, followed by Random Forest and clustering techniques. Advanced or specialized approaches such as fuzzy models, neural networks, feature selection, CRISP-DM process frameworks, similarity modeling (Jaccard), and KNN appear less frequently, highlighting opportunities for richer comparative studies and integrated frameworks in future EDM research.

5. Conclusion

This study presented a comparative examination of data mining techniques applied to student performance analysis in education, drawing insights from recent empirical studies and systematic reviews. The analysis shows that traditional machine learning models, particularly decision trees and random forest, are the most widely used due to their strong predictive capabilities and adaptability to educational data. Ensemble models consistently demonstrate high accuracy, while simpler algorithms such as naïve Bayes and regression models offer advantages in interpretability and ease of implementation. The review also highlights the growing use of clustering, feature selection, and process-oriented frameworks to enhance analytical effectiveness. Despite these strengths, several limitations persist, including reliance on context-specific datasets, challenges related to data quality and imbalance, and limited application of advanced models in school-level education. Moreover, the lack of standardized evaluation frameworks and insufficient focus on explainability reduce the practical impact of many proposed solutions. Addressing these challenges is essential for translating EDM research into meaningful educational improvements. Future work should focus on developing scalable, explainable, and cross-context data mining frameworks validated on diverse and real-world educational datasets.

References

1. Ordonez-Avila, Ricardo, Nelson Salgado Reyes, Jaime Meza, and Sebastián Ventura. "Data mining techniques for predicting teacher evaluation in higher education: A systematic literature review." *Heliyon* 9, no. 3 (2023).
2. Ashish, L., and G. Anitha. "Transforming Education: The Impact Of ICT And Data Mining On Student Outcomes." In *2024 7th International Conference on Circuit Power and Computing Technologies (ICCPCT)*, vol. 1, pp. 675-683. IEEE, 2024.
3. Staneviciene, Evelina, Daina Gudoniene, Vytenis Punys, and Arturas Kukstys. "A case study on the data mining-based prediction of students' performance for effective and sustainable e-learning." Sustainability. 16, no. 23 (2024): 1-15.
4. Kaur, Kanksha, and Omdev Dahiya. "Role of educational data mining and learning analytics techniques used for predictive modeling." In 2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM), pp. 1-6. IEEE, 2023.
5. Missah, Yaw Marfo, Fuseini Inusah, Najim Ussiph, and Twum Frimpong. "Systematic Review of Data Mining in Education on the Levels and Aspects of Education." (2023).
6. Sun, X., 2024. Educational Technology Based on Data Mining: Mining Student Behavior Patterns to Optimize Teaching Strategies. International Journal of High Speed Electronics and Systems, p.2540101.
7. Abideen, Z.U., Mazhar, T., Razzaq, A., Haq, I., Ullah, I., Alasmary, H. and Mohamed, H.G., 2023. Analysis of enrollment criteria in secondary schools using machine learning and data mining approach. *Electronics*, *12*(3), p.694.
8. Wang, Jinhui, Jiande Sun, Ran Lu, Yawen Chen, Cheng Su, and Zhen Zhao. "A Practical Study on Statistical Analysis of Secondary School Achievement and Case Teaching Based on Data Mining." In *2024 14th International Conference on Information Technology in Medicine and Education (ITME)*, pp. 498-502. IEEE, 2024.
9. Syed Mustapha, S. M. F. D. "Predictive analysis of students' learning performance using data mining techniques: A comparative study of feature selection methods." *Applied System Innovation* 6, no. 5 (2023): 86.
10. Alghamdi, Amnah Saeed, and Atta Rahman. "Data mining approach to predict success of secondary school students: A Saudi Arabian case study." *Education Sciences* 13, no. 3 (2023): 293.

11. Chan, Ka Ian, Philip IS Lei, and Patrick Cheong-Iao Pang. "A literature review on educational data mining with secondary school data." In *Proceedings of the 9th International Conference on Education and Training Technologies*, pp. 1-7. 2023.
12. Collier, Z., Sukumar, J. and Barmaki, R., 2024. Discovering educational data mining: An introduction. *Practical Assessment, Research, and Evaluation*, *29*(1).
13. Assiri, Basem, Mohammed Bashraheel, and Ala Alsuri. "Enhanced student admission procedures at universities using data mining and machine learning techniques." *Applied Sciences* 14, no. 3 (2024): 1109.
14. Huerta, C.M., Atahua, A.S., Guerrero, J.V. and Andrade-Arenas, L., 2023. Data mining: Application of digital marketing in education. *Advances in Mobile Learning Educational Research*, *3*(1), pp.621-629.